

THE UNIVERSITY OF TEXAS
AT AUSTIN

WHAT STARTS HERE CHANGES THE WORLD

CHE384, From Data to Decisions: Measurement, Uncertainty, Analysis, and Modeling

Lecture 53

Principal Component Analysis

Chris A. Mack
Adjunct Associate Professor

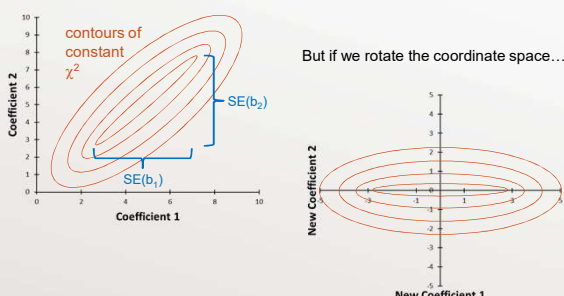
<http://www.lithoguru.com/scientist/statistics/>

© Chris Mack, 2016 Data to Decisions 1

THE UNIVERSITY OF TEXAS
AT AUSTIN

WHAT STARTS HERE CHANGES THE WORLD

When Two Predictors are Correlated



But if we rotate the coordinate space...

© Chris Mack, 2016 Data to Decisions 2

THE UNIVERSITY OF TEXAS
AT AUSTIN

WHAT STARTS HERE CHANGES THE WORLD

Linear Constraints

- Multicollinearity often comes about because of constraints between predictor variables
 - e.g., $x_2 - x_4 + x_5 = 5$, $x_6 = 0.8x_7$, etc.
- Theoretical considerations can often be used to identify constraints; if so, use them to simplify the regression model (and reduce multicollinearity)
- Principal Component Analysis can help to identify **unknown linear constraints**
 - But not all constraints are linear (e.g., $x_2x_4 = 5$)

© Chris Mack, 2016 Data to Decisions 3

THE UNIVERSITY OF TEXAS
AT AUSTIN

WHAT STARTS HERE CHANGES THE WORLD

Principal Component Analysis

- Rotate coordinate systems to create an orthogonal set of new regressor variables
 - Each new variable (called a principal component) is a linear combination of the original variables
- Steps to do this:
 - Calculate the correlation matrix of predictors
 - Find the eigenvalues and corresponding eigenvectors of the correlation matrix
 - Orthogonalize the design matrix by multiplying by a rotation matrix made up of the eigenvectors

© Chris Mack, 2016 Data to Decisions 4

THE UNIVERSITY OF TEXAS
AT AUSTIN

WHAT STARTS HERE CHANGES THE WORLD

Eigenvectors and Eigenvalues

- For a square $p \times p$ matrix A , the **eigenvectors** v and **eigenvalues** λ are defined by $Av = \lambda v$
- The eigenvalues are found by the p -roots of the equation $|A - \lambda I| = 0$
- Given the eigenvalues, the eigenvectors (one for every eigenvalue) can be determined by substitution into the original equation

© Chris Mack, 2016 Data to Decisions 5

THE UNIVERSITY OF TEXAS
AT AUSTIN

WHAT STARTS HERE CHANGES THE WORLD

Orthogonalizing the Design Matrix

- Create a $p \times p$ rotation matrix T with each column equal to one eigenvector from correlation matrix
 - Order the columns by eigenvalue from large to small
- Create a "rotated" coordinate design matrix as $Z = XT$
 - Each column of Z is a "principal component", orthogonal to all the other principal components

$$z_1 = t_{11}x_1 + t_{12}x_2 + \dots + t_{1p}x_p$$

$$z_2 = t_{21}x_1 + t_{22}x_2 + \dots + t_{2p}x_p$$

etc.

Each new variable z_k is a linear combination of old variables x_i

© Chris Mack, 2016 Data to Decisions 6

THE UNIVERSITY OF TEXAS AT AUSTIN WHAT STARTS HERE CHANGES THE WORLD

Properties of the Rotated Design

- The “rotated” design matrix is $Z = XT$
- Each column of Z is a “principal component” (PC), orthogonal to all the other PCs
- The first PC (largest eigenvalue) accounts for the maximum amount of variance in the predictors
- Each PC has mean = 0, and variance equal to the eigenvalue
- For a small eigenvalue, the PC is about constant at a value of 0; this is a constraint on the predictors that can be used to simplify the model
 - $z_k = t_{k1}x_1 + t_{k2}x_2 + \dots + t_{kp}x_p \approx 0$

© Chris Mack, 2016 Data to Decisions 7

THE UNIVERSITY OF TEXAS AT AUSTIN WHAT STARTS HERE CHANGES THE WORLD

Principal Component Analysis

- What do the Principal Components mean?
 - Sometimes, we don't care – we just want a simple model with predictive ability
 - But sometimes, the principal components might be revealing!
 - Why are the principal components this specific linear combination of other variables?
 - What constraints (small eigenvalues) are revealed?
 - Look at the correlation matrix between the PCs and the original predictor variables

© Chris Mack, 2016 Data to Decisions 8

THE UNIVERSITY OF TEXAS AT AUSTIN WHAT STARTS HERE CHANGES THE WORLD

Principal Component Regression

- We can create a regression model using the Z design matrix rather than X
- Frequently, some of the principal components will not be correlated with the response
 - They can be excluded from the model
- Since each PC is orthogonal to the others, a simple t-test of model coefficients can be used to keep or exclude model terms
 - There is no multicollinearity

© Chris Mack, 2016 Data to Decisions 9

THE UNIVERSITY OF TEXAS AT AUSTIN WHAT STARTS HERE CHANGES THE WORLD

PCA Example: Two Regressors

- Admittedly, this is a case where we would probably never use PCA (since there are two few regressors), but it will illustrate the methods
- Recall: $\bar{X}^T \bar{X} = \begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix}$ $\lambda_1 = 1 + r_{12}$
 $\lambda_2 = 1 - r_{12}$
- Find the eigenvectors:

$$\begin{bmatrix} 1 - \lambda_j & r_{12} \\ r_{12} & 1 - \lambda_j \end{bmatrix} \begin{bmatrix} v_{j1} \\ v_{j2} \end{bmatrix} = 0 \rightarrow \begin{cases} (1 - \lambda_j)v_{j1} + r_{12}v_{j2} = 0 \\ r_{12}v_{j1} + (1 - \lambda_j)v_{j2} = 0 \end{cases}$$
- Plugging in λ_1 and then λ_2 , we find the two eigenvectors v_1 and v_2 .

$$v_1 = k_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad v_2 = k_2 \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$
 where k_1 and k_2 are arbitrary constants, often chosen to make the vectors unit length

© Chris Mack, 2016 Data to Decisions 10

THE UNIVERSITY OF TEXAS AT AUSTIN WHAT STARTS HERE CHANGES THE WORLD

PCA Example: Two Regressors

- The rotation matrix is formed by putting the eigenvectors into columns:

$$T = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \quad \text{where } k_1 = k_2 = 1/\sqrt{2}$$
- The rotated design matrix becomes $Z = \bar{X}T$. Carrying out the multiplication, the new rotated variables are

$$z_1 = 1/\sqrt{2}(\tilde{x}_1 + \tilde{x}_2)$$

$$z_2 = 1/\sqrt{2}(\tilde{x}_1 - \tilde{x}_2)$$
- Now regress y on these two rotated variables
 - Note that z_1 and z_2 are orthogonal
 - Only for the special case of two regressors is the rotation matrix independent of the data (that is, r_{12} does not appear in T or in z_1 and z_2).

© Chris Mack, 2016 Data to Decisions 11

THE UNIVERSITY OF TEXAS AT AUSTIN WHAT STARTS HERE CHANGES THE WORLD

Lecture 53: What have we learned?

- Explain how constraints between regressor variables lead to multicollinearity
- Why is PCA sometimes described as a rotation of the parameter space?
- What does a small eigenvalue tell you about that principal component?
- How can PCA be used to improve regression?

© Chris Mack, 2016 Data to Decisions 12